Automatic Acquisition of Class-based Rules for Word Alignment

Sur-Jin Ker and Jason J.S. Chang
Department of Computer Science
National Tsing Hua Universitv
Hsinchu, Taiwan

Abstract

In this paper, we describe an algorithm for aligning words with their translation in a bilingual corpus. Existing algorithms require enormous bilingual data to train statistical word-to-word translation models. Using word-based approach, frequent words with consistent translation can be aligned at a high precision rate. However, less frequent words or words with diverse translations usually do not have statistically significant evidence for confident alignment. Incomplete or incorrect alignments consequently result. Our algorithm attempts to handle the problem using a hierarchical class-based approximation of translation probabilities. The translation probabilities are estimated using class-based models on 3 levels of specificity. We found that the algorithm can provide translation probability for more word pairs at the cost of slightly lower degree of precision, even when a small corpus was used in training. We have achieved an application rate of 81.8% and precision rate of 93.3%. The algorithm also offer the advantage of producing word-sense disambiguation information.

1. Introduction

Much of the recent interests in bilingual corpora was initiated by Brown et al. (1990). They advocated a new approach to machine translation in which the bilingual corpora are *aligned* to reveal the mapping between text in one language and its translation in another language. This mapping is formally represented as statistical machine translation model. This model can be understood as a word by word model for generating (*language model*) of simple sentences S and translation (*translation model*) S to a sentence T in another language. The i-th word(s) s in S is considered to *connect* with its translation, the j-th word(s) t in T. Under the model, the probability of the connection (s, t) can be obtained by considering three aspects: lexical translation (t(t | s), the relation of s producing t in translation), fertility (f($|t|$||$s|$), the relation of change in number of words), and distortion (dis(i | j, l, m), the relation between the position i, j of s, t and the respectively length l, m of S, T). Probabilities are associated with these three relations of connection. A bilingual corpus annotated with these connection information can be utilized to estimate the parameters in a statistical translation model. Subsequently, the model can be used together with a bigram or trigram language model in machine translation.

We present a rule-based algorithm for word alignment. We refer to this algorithm as SenseAlign. It relies on an automatic procedure for the acquisition of class-based rules. It does not employ word by word translation probabilities to identify alignment; nor does it use a lengthy iterative EM algorithm for finding such probabilities. The algorithm attempts to handle the problem of undersampling by approximating word-to-word translation probability using hierarchical classifications of words. The translation probabilities are estimated using class-based rules on different levels of specificity. We found that the algorithm can provide translation probability for more source-target word pairs at the cost of slightly lower degree of precision. That allows the algorithm to work in situation where only a small bilingual corpus is available. We have achieved an application rate of 81.8% and precision rate of 93.3%. Since the rules are all based on word sense distinction, word sense ambiguity is also resolved in the process of alignment.

The paper is organized as follows. In the next section, we describe SenseAlign and discuss its main components. We give the results of inside and outside tests in Section 3. In Section 4, we compare SenseAlign to several other approaches that have been proposed in computational linguistic literature. Finally, in Section 5, we consider ways in which our present methods might be extended and improved.

2. The Word Alignment Algorithm

SenseAlign is a word alignment system that utilizes both existing and acquired lexical knowledge. The system contains the following components and distinctive features.

2.1 Two lexical preprocessors

Morphological analysis, part-of-speech tagging, idioms identification are performed for the two languages involved. In addition, certain morpho-syntactic analyses are performed to handle structures that are specific only to one of the two languages involved. By doing so, we bring the sentences closer to each other in number of words. Furthermore, since sense ambiguity are often related to POS ambiguity, resolution of POS ambiguities reduces the degree of sense ambiguity. These two factors make the alignment task a lot easier.

The part-of-speech taggers for the two languages involved are trained using a strategy proposed by Brill (1993). We use the tag set in Brown Corpus for the English tagger and adopt a part-of-speech system proposed by Chao (1968). See Table 1.1 and 1.2 for the two tag sets.

2.2 Two thesauri for classifying words

Classification is intended to allow a word to align with a target word using the collective translation tendency of words in the same class. Class by class translation probabilities obviously have much less parameters and are easier to estimate. In this work, the categories for Chinese text are taken from a thesaurus for Mandarin Chinese (Mei 1984, CILIN henceforth). The categories for English text are taken from the Longman Lexicon (McArthur 1992, LEXICON henceforth).

The categories in CILIN are organized as a conceptual ontology of 3 levels: gross categories, intermediate categories and detail categories. Each gross category has from 1 to 19 intermediate categories listed under it and there are 94 of intermediate categories in total. Under each intermediate category, there are from 5 to 55 detailed categories and 2 digits are used to represent detailed categories. There are 1428 detailed categories.

Unlike CILIN, the category of LEXICON is organized mainly according to subject matter. On the first level, there are 14 major subjects denoted with reference letters from "A" to "N". There are from 7 to 12 titles listed under a subject. Under each title, there are from 10 to 50 *sets* of related words. Each set is given a 3-digit reference number. The titles are not reflected in the original LEXICON reference code. In order to represent this implicit grouping, we assign a lower case letter to each title. For example, "*objects generally*" is denoted using the letter *b* and the reference code *H030* is replaced with *Hb030*. So each set is denoted by a upper case SUBJECT letter, a lower case TITLE letter and a 3-digit SET number. There are 2504 sets in total.

2.3 An learner of class-based rules

This procedure employs a greedy method to find rules that can provide optimal alignments in a corpus of bilingual sentences. The rules that can provide the most instances of plausible connection is selected first. These connections that are applicable for the selected rule are then removed. This process iterates until certain low threshold on the number of applicable instances. This contrast with mutual-information-based approach that find statistical significant word pairs that often have limited presence in the data.

SenseAlign uses class-based alignment rules as the main mechanism. These rules are

represented in the form of a pair of class codes in the two languages. We present our algorithm for learning such rules automatically from examples. This procedure employs the greedy method to find set of rules that can provide optimal alignments in a corpus of bilingual sentences.

Table 1.1. The part-of-speeches used in the English tagger

POS	Meaning	Examples	POS	Meaning	Examples
AB	pre-qualifier		NP	proper noun	John, Mary
AT	Article	a, the, no	NR	adverbial noun	home, west, tomorrow
BE	be	be	PN	nominal pronoun	everybody, nothing
CC	coordinating conj.	and, or	PP	personal pronoun	I, you, he
CD	cardinal numeral	two , 2	RB	adverb	well
CS	subordinate conj.	if, although	RN	nominal adverb	here, then
DT	determiner	this, that, some, any	RP	adverb or particle	up, down
IN	preposition	in, at	TO	infinite marker	to
JJ	adjective	good, bad	UH	interjection	oh
NN	noun	boy, girl	VB	verb	make, get

Table 1.2. The part-of-speeches used in the Chinese tagger

POS	Meaning	POS	Meaning
A	Adjective	Na	Common noun
C	Conjunction	Nb	Proper noun (person)
Da	Quantity adverb	Nc	Proper noun (locative)
Db	Judgment adverb	Nd	Temporal
Dc	Negation adverb	Ne	Determinant or quantifier
Dd	Temporal adverb	Nf	Measure noun
De	Degree adverb	Ng	Locative noun
Df	Locative adverb	P	Preposition
Di	Aspect	V	Verb
Dj	Question adverb	VH	State Verb

The rules that can provide the largest number of instances of plausible alignment is preferred and selected first. At this stage we have no information of what class of words in one language is likely to align some classes of words in the other language. We match up randomly classes of words with compatible POS across the two sentences to form tentative *alignment rules*. When applying one of such rules to a sentence, we use the term *fan-out* to denote the number of words that match the two classes. For example, when applying a tentative alignment rule (C, D) to a sentence pair (S, T), we say that (C, D) have a fan-out of (n, m) if there are n and m words in S and T that belongs to the classes C and D respectively. For a alignment of fan-out (n, m), we define the *degree* of fan-out of a connection c as $f_c = n \times m$.

After producing tentative alignment rules for all the sentences, we make a conservative estimation of *applicability* by counting the number of connections where a rule is applicable with a fan-out of (1-1). The rule with the highest *estimated applicability* is selected. Sentences where the rule applies are identified and applicable connection are removed from the sentences. The connections that are inconsistent with the applicable connections are also removed. This process iterates until the highest applicability go below a certain threshold.

This procedure of learning alignment rules is applied to the detailed categories of CILIN and sets of LEXICON to produce 392 rules. See Table 2 for the 10 rules with the highest applicability. Clearly, 392 rules do not cover all English words in the 2504 class, nor all Chinese words in the 1436 classes. To remedy this, we repeat the procedure for broader 2-letter classes represented by 130 titles in LEXICON and 205 intermediate categories in CILIN. Similarly, a alignment rules are also learned for one-letter classes of 15 subjects in LEXICON and 23 gross categories in CILIN.

See Table 3 for the highest scored rules on the 2-letter level.

Table 2. Ten rules with the highest applicability

Rule#	Appl.	POS		Rule	Gloss for LEXICON	Gloss for CILIN
1	642	VB	V	Ma001, Hj63	moving, coming, and going	來(lai, come), 去(qu, go)
2	459	NN	Na	Jh210, Di19	jobs, trade and professions	職業(zhie, job)
3	440	NN	Na	Md108, Bo21	trains	車(che, car)
4	418	JJ	A	Lg202, Eb28	new	新(xin, new), 新鮮(xinxian, fresh)
5	367	NN	Na	Da003, Bn01	things built and lived in	建築(jianzhu, building)
6	362	VB	V	Gc060, Hi16	speaking, and telling	介紹(jieshao, introduce)
7	349	JJ	A	Fc050, Ed03	the right qualities	好(hao, good), 壞 (huai, bad)
8	310	NN	Na	Lh226, Tl18	measuring time	年(nian, year)
9	303	NN	Na	Ca002, Ab04	man and woman	嬰兒(ienger, baby)
10	302	VB	V	Fb020, Gb09	liking and loving	喜歡(xihuan, like), 愛(ai, love)

Table 3. Three rules with the highest estimated applicability on 2-letter level

Rule#	Appl.	POS		Rule	Gloss for LEXICON	Gloss for CILIN
1	1628	VB	V	Ma, Hj	moving, coming, and going	生活(shenghuo, life)
2	1251	VB	V	Gc, Hi	communicating	社交(shejiao, socialize)
3	980	JJ	A	Mh, Ed	locating and direction	性質(xingzh, property)

2.4 Relative distortion

As pointed out in Dagan, Church and Gale (1992) the distribution function for distortion, Pr($i \mid j, l, m$) has too many parameters to estimate reliably. In other words, it is likely that Pr($i \mid j, l, m$) is very close to Pr(i+1 $\mid$ j+1, l, m). Since the translation process tends to preserve contiguous structures, parameters in an model of distortion based on absolute position are highly redundant. Therefore, it is advantageous to replace probabilities of the form Pr($i \mid j, l, m$) with a smaller set of distribution probabilities for *relative distortion*.

Assuming some alignments have been established before we evaluate the alignment (i, j) with respect to its distortion. The closest words on both sides of the i-th word with an established alignment are i_L and i_R respectively. The established connection for position i_L and i_R are j_L and j_R respectively. Relative distortion $rd(i, j)$ is approximated using the following formula:

$$d_L = (j - j_L) - (i - i_L)$$
$$d_R = (i - i_R) - (j - j_R)$$
$$rd(i, j) = \min(|d_L|, |d_R|)$$

2.5 Similarity between connection and dictionary translations

Using a bilingual dictionary directly for word alignment is surprising ineffective; only 16% of correct connections are listed in a common bilingual dictionary. However, in 40% of correct connections, the translation in the connection and dictionary translation have at least one Chinese character in common. In other words, these translations can be considered as synonyms that would appear in a thesaurus. In order to take advantage of this thesaury effect in translation, some means of quantifying the degree of similarity between words are required. A variety of distance and similarity measures is given by Anderberg (1973). For simplicity, we select the Dice coefficient to calculate the similarity between of connection and dictionary translation. The equation of Dice coefficient is shown as equation (1).

(1) $$\frac{2|E|}{|C| + |D|}$$ where C = Chinese translation in connection

D = Chinese characters in dictionary

E = common Chinese characters in C and D

2.6 A procedure for evaluating the probability of connection candidates

The evaluation is based on composite scores of applicability, specificity, fan-out of class-based alignment rules, relative distortion probabilities, and evidence from bilingual dictionaries.

Obviously, an alignment rule such as (*Ma001*, *Hj63*) is very *specific* since it only applies to a small set of words: 'move,' 'come,' 'go,' 'pass,' 'get out,' 'set out,' '來 (lai),' '去 (qu),' etc. The more specific a rule is, the more reliable are the alignments it predicts. Therefore, we need to have the following definition for *specificity s* for alignment rule *r*:

s_r **= 1/(the number of word pairs for which *r* is applicable) .**

For example, there are 27 English words listed under *Ma001* and 48 Chinese words listed under *Hj63*, so the specificity of (*Ma001*, *Hj63*) is 1/1296.

An alignment rule such as (*Ma001*, *Hj63*) also has high degree of *applicability* for it applies to many instances of word pairs in the bilingual corpus. The higher applicability an alignment rule has, the more reliable are the alignments it predicts. Therefore, we define *applicability* of an alignment rule *r* as follows:

a_r **= (# of word pairs for which *r* is applicable) / (# of bilingual sentences) .**

2.7 A decision procedure for selecting the preferred connection

A greedy procedure is employed to determine the alignment. The connection with highest composite score is selected first. The candidates that are inconsistent with the selected connection are then removed from the list. The procedure of selecting connections iterates until no candidate are left in the list. Note that relative distortions are evaluated in each iteration with respect to connections that have been selected.

Fan-out is the numbers of source and target words that can be aligned using alignment rules. The smaller fan-out is, the more reliable the alignment is. An alignment with fan-out 1-1 receives a weight of 100 points. An connection candidate (s, t) receives $100/(i \times j)$ points if it is applicable for an alignment rule of fan-out (i, j).

Relative Distortion measures the distance of alignment site relative to those of neighboring words. Obviously, connection candidate with small distortion should be preferred. Therefore for the distortion factor, we give each connection candidate a weight inversely proportional to square of its relative distortion. One examples are given to show how distortion play a role in determining the correct connection.

(1e) Please answer all **questions** on this **list**.

(1c)	請	回答	本	表	上	之	所有	問題
	qing	huida	ben	**biao**	shang	zh	suoiou	**wuti**
	Please	answer	this	**list**	on	CTM	all	**question**

In example 1, the initial values of relative distortion for connection candidates are shown in table 4. Note that a lot of correct connections receive a value of 3 for their relative distortion. These large values of distortion are due to the transfer of prepositional phrase "on this list" forward to the front of the attached noun phrase, "all questions." After the connection (question, 問題 wuti) is selected, the relative distortion of the word "all" is reduced to 0. See Table 4 and Table 5 for details.

Similarity between connection and dictionary translations. It is conceivable that a sentence mentions words (such as "yesterday" and "today") that belong to the same class (Lh225). In such

event, class-based rule create alignment ambiguity. Simple dictionary lookup can resolve such kind of ambiguity. The connection candidates that are confirmed by dictionary translations received a weight of 100.

Specificity is the number of word pairs that a rule can conceivably apply to. The more specific a rule is, the smaller is the set of applicable word pairs. The smaller the set is, the more likely it contains truly interchangeable translations. Therefore, an connection is given a weight according to the specificity of the rule being applied to. In the following example, the connection (know, 知道 zhidao) is preferred because it is predicted by a more specific rule (Gb030, Gb08) over an incorrect connection (cat, 狗 gou) predicted using a less specific rule (Ac054, Bi08).

(2e) I only knew that it is the dog not the cat that bit me.

(2c)

我	只	知道	咬	我	的	是	狗	不是	貓。
wuo	zhi	zhidao	yiao	wuo	de	shi	gou	bushi	mao.
I	only	know	bit	I	DE	BE	dog	NOT	cat.

English Word	English Code	Spec 1	Chinese Word	Chinese Code	Spec 2	Spec 1 * Spec 2	Appl.
know	Gb030	5	知道	Gb08	131	655	234
dog	Ac054	50	狗	Dd15	140	7000	144
cat	Ac053	33	貓	Bi08	42	1386	55

Table 4. The Relative Distortion (*rd*) of word pairs (Initially)

English	Position	POS	Chinese	Position	POS	d_L	d_R	*rd*
answer	2	VB	回答 (huida, answer)	2	V	0	1	0
all	3	AT	所有 (suoyou, all)	7	Ne	4	-3	3
question	4	NN	表 (biao, list)	4	Na	0	1	0
question	**4**	**NN**	**問題 (wunti, question)**	**8**	**Na**	**4**	**-3**	**3**
on	5	IN	上 (shang, up)	5	Ng	0	1	0
this	6	AT	本 (ben, this)	3	Ne	-3	4	3
list	**7**	**NN**	**表 (biao, list)**	**4**	**Na**	**-3**	**4**	**3**
list	7	NN	問題 (wunti, question)	8	Na	1	0	0

Table 5. The Relative Distortion (*rd*) after initial selection of (question, 問題)

English	Position	POS	Chinese	Position	POS	d_L	d_R	*rd*
answer	2	VB	回答 (huida, answer)	2	V	0	1	0
all	**3**	**AT**	**所有 (suoyou, all)**	**7**	**Ne**	**4**	**3**	**0**
on	5	IN	上 (shang, up)	5	Ng	0	-3	0
this	6	AT	本 (ben, this)	3	Ne	-3	4	3
list	7	NN	表 (biao, list)	4	Na	-3	4	3

Applicability is the number of instances of word pairs that contribute to the acquisition of the rule. Higher applicability means more word pairs are found in the training phase to support the rule. Therefore, a connection candidate predicted by a rule with higher applicability should receive a higher weight. This weighting scheme also result in correct alignment of more candidates since the rule that can be applied to more instances is preferred. Connection candidate receiving more weight for the applicability factor are (know, 知道 zhidao) and (dog, 狗 gou) in example 2.

2.8 Alignment algorithm

Our algorithm for word alignment is a decision procedure for selecting the preferred connection

from a list of candidates. Two dummies are placed to the left (right) of the first (last) words of the source and target sentences. The left dummy in the source and target sentences connect with each other. Similarly, the right dummies connect with each other. The initial list of selected connection contains these two connection of dummies. This establish initial anchor points for calculate relative distortion. The highest scored candidate is selected and added to the list of solution. The newly added alignment serves as additional anchor for more accurate estimation of relative distortion. The connection candidates that are inconsistent with the selected connection are also removed from the list. Subsequently, the rest of the candidate that is evaluated again. The algorithm of SenseAlign are given as Figure 1. We summarized all factors in Table 6, and the weight of each factor are shown in Table 7.

Table 6. Summary of factors and formula used in SenseAlign

Fan-out : $f_c(C, D) = n \times m$	where n = the number of C-class words in S and m = the number of D-class words in T
Specificity : $S_r(e,c) = \frac{1}{W_r(e,c)}$	
$W_r(e, c) = W_e \times W_c$	where W_e = the number of English word in class e, W_c = the number of Chinese word in class c
Applicability : $a_r = \frac{C_r}{B}$	where C_r = the number of connections for which r in corpus, and B = the number of bilingual sentence in corpus
Relative Distortion : $rd(i,j) = \min(\lvert d_L \rvert, \lvert d_R \rvert)$	
$d_L = (j - j_L) - (i - i_L)$	where i = the subscript of source sentence
$d_R = (i - i_R) - (j - j_R)$	j = the subscript of target sentence
	i_L (i_R) = the closest words on left (right) side of the i-th word with an established alignment,
	j_L (j_R) = the established alignment for i_L (i_R).
Dice coefficient : $Sim = \frac{2 \lvert E \rvert}{\lvert C \rvert + \lvert D \rvert}$	where C = Chinese translation in connection
	D = Chinese characters in dictionary
	E = common Chinese characters in C and D

3. Experiments with SenseAlign

In this section, we show the results of various algorithm for word alignment. We use the 25,000 bilingual examples (English-Chinese sentence pairs) of Contemporary English (Longman 1992) as the training data. Our training data was primarily used to selected the aligning rule by a greedy learner. The performance of the algorithm was then tested on the two set of inside and outside data. The inside test use 50 sentence pairs from LecDOCE as input. The input data for outside test are 416 sentence pairs from a book on English sentence patterns.

The first experiment is designed to show the performance of an naive algorithm (DictAlign) based on bilingual dictionary. We have found that although DictAlign produce high precision alignment. The applicability is below 16%. However, the applicability can be increased greatly, if one take advantage of the thesaurus effect in the character of the target language. In our case of using thesaury effect, the applicability can be increase almost 3 folds to 40%, at the expense of 10% decrease in precision. See Table 8 for details.

Table 7. Factor types with weights of applying r to connection c

Factor type	Weight
Fan-out	$100/f_c$
Relative distortion emphasis	$10/(1+rd_c^2)$
Specificity emphasis	$10 \times s_r$
Dictionary evidence emphasis	$100 \times Sim$
Applicability emphasis	$100 \times a_r$

1. Read a pair of English-Chinese sentences.
2. Two dummies are replace to the left of the first and to the right of the last word of the source sentence. Similar two dummies are added to the target sentence. The left dummy in the source and target sentences align with each other. Similarly, the right dummies align each other. This establish anchor points for calculate relative distortion score.
3. Perform the part-of-speech tagging and analysis for sentence in both languages.
4. Lookup in LEXICON and CILIN to determine the classes consistent with the part-of-speech analyses.
5. Calculate a weight for each connection candidate according to fan-out, applicability, specificity of alignment rules, relative distortion, dictionary evidence.
6. The highest scored candidate is selected and added to the list of alignment.
7. The connection candidates that are inconsistent with the selected connection are also removed from the candidate list.
8. The rest of the candidate that is evaluated again according to the new list of connection.
9. The procedure iterates until all words in the source sentence are aligned.

Figure 1. Alignment Algorithm of SenseAlign

Table 8. The performance of DictAlign

	Inside Testing				Outside Testing			
Dice Coefficients	Mapped No.	Correct No.	Appl.	Precision	Mapped No.	Correct No.	Appl.	Precision
1.0	59	56	15.3%	94.9%	499	486	16.8%	97.4%
0.67	113	100	29.4%	88.5%	970	865	32.7%	89.2%
0.5	151	124	39.2%	82.1%	1221	1046	41.1%	85.7%

In our second experiment, we use SenAlign described above for word alignment except that no bilingual dictionary is used. The result is shown in Table 9. In our third experiment, we use full SenAlign to aligning our testing data. The results in Table 10 show that acquired lexical information augmented and existing lexical information such as a bilingual dictionary can supplement each other to produce best alignment results.

4. Previous Works

We will briefly compare our algorithm with several other approaches to word alignment and sense disambiguation that have been suggested.

Table 9. The results of SenseAlign

Inside Testing				Outside Testing			
Mapped No.	Correct No.	Appl.	Precision	Mapped No.	Correct No.	Appl.	Precision
237	213	61.7%	89.9%	1913	1721	66.8%	90.0%

Table 10. The results of Full SenseAlign

Inside Testing				Outside Testing			
Mapped No.	Correct No.	Appl.	Precision	Mapped No.	Correct No.	Appl.	Precision
314	293	81.8%	93.3%	2424	2265	84.7	93.4%

4.1 Word-based EM Algorithms for Word Alignment

In the context of statistical machine translation, Brown et al. regard translation to be a process of recovering from a noisy channel that maps a target sentence T to a source sentence S with probability Pr(S | T). To translate a sentence SS in the source language amounts to finding a target sentence TS to optimize the probability, Pr(S=SS | T=TS). The probability Pr(S | T) is computed using the concept of alignment, which is a set of connections between a word (or words) in S and a word (or words) in T. In their paper, Brown et al. present a series of 5 models for Pr(S | T). The first two models have been used in research on word alignment.

Model 1 assumes that Pr(S | T) depends only on the probability that the i-th word SW in S is translated to the j-th word TW in T. Such a translation pair (SW, TW) is called a connection with a probability of t(TW | SW). Model 2 improves on Model 1 by considering the role that the positions (i, j) of the connection play in Pr(S |T).

Brown et al. report that the first experiment on alignment of the Hansards on the levels of both sentences and words using an EM algorithm. The trained model produced 17 acceptable translation for 26 testing sentences. However, the degree of success in word alignment was not reported.

Dagan, Church and Gale (1992) observed that noisy bilingual text is difficult to distinguish sentence boundaries reliably. So they proposed to align words directly without the preprocessing phase of sentence alignment. They was shown that 60.5% of 65,000 words are aligned correctly. For 84% of the words, the offset from the correct alignment was at most 3. Chen and Chang (1994) proposed using part-of-speech information and position for word alignment. The main idea is to trade some precision for higher applicability, smaller model and faster training.

4.2 Algorithms based on word by word mutual information

Gale and Church (1991) proposed a word alignment technique, where the t(SW | TW) is estimated indirectly. For pairs of a source word SW and a target word TW, the correlation of the translation of SW and TW is estimated using a χ^2-like statistics, based on a two by two contingency table. A list L of 13,466 highly correlated word pairs are selected. They reported 60% applicability and 95% precision rate on 800 English-French sentences. The use of contingency table differs with the EM algorithm in that the co-non-occurrence is utilized.

4.3 Hybrid Models for Word Alignment and Sense Ambiguity

This approach seeks to combine a variety of morphological, syntactic, semantic, positional factors for ranking connection candidate in word alignment. There are two ways to go about this.

One way is to use some of these factors as filters in preprocessing to reduce ambiguity and decrease the number of candidates. Brown et al. (1992) suggests that part of speech tagging and

normalization of syntactical structures such as adjective-noun and adverbs for both languages involved to narrow down the differences.

Another way is to treat all factors uniformly using a weighting scheme. On this view, the score of a candidate is a composite of several distinct factors, each of which reflects the prominence of the candidate with respect to a specific type of information or property. Our algorithm also used a mixed evaluation strategy. We have taken inspiration from the discussions of the χ^2-like statistical procedures and the fan-out factor in the work cited above, but we try to avoid the disadvantage of undersampling. Specifically, we have choose to work with class-based rules which are acquired to maximize both applicability and precision. In general, it seems to us that it is possible to trade small percentage of precision to gain a substantial increase in applicability. Our results suggest that mixed strategies can yield a broad coverage and high precision word alignment and sense resolution system which can produce rich information for MT. Moreover, the word sense information can provide a certain degree of generality which is lacking in most statistical procedures.

4.4. Works related to Word-alignment

In the methods described above, only one-to-one translation probabilities between source and target words are considered. Kupiec (1992) described an experiment on alignment of noun phrases, while Eijk (1994) and Daille (1994) have done similar works focusing on acquisition of bilingual terminology.

5. Concluding remarks

This paper has presented an algorithm which is effective in identifying words and their translation in a bilingual corpus. It is effective for specific linguistic reasons. The great majority of words in bilingual sentences have diverging translation and these translations are not often found in bilingual dictionary. However, these deviation are largely limited within the classes defined by thesauri. Therefore, by using a class-based approach, the complexity of the problem can reduced in the sense that less number of candidates need to be considered with greater chance of finding the correct translation. The complexity can be further reduced by exploiting the fact that most syntactic structures are preserved across translation. Using a distortion model that measure positional deviation across translation relative the adjacent words, complexity can be further reduced by a better estimation of the probability of a translation.

The performance of the algorithm discussed here can surely be improved by the enhancement in the various components of the algorithm, such as morphological analyses, bilingual dictionary, monolingual thesauri, rule acquisition. However, what we have presented here is a workable core for processing bilingual corpus. The algorithm can produce good word-alignment results with sense tagging which can provide a basis for such NLP tasks as word sense disambiguation and PP attachment.

While this paper has specifically addressed only English-Chinese corpus, the linguistic issues that motivated the algorithm are quite general and are to a great degree language independent. If that is truly the case, the algorithm presented here should be adaptable to other language pairs. The prospects for Japanese in particular seem promising, for example, Matsumoto et al. (1993) have already implemented an structural alignment system for English-Japanese sentences using a Japanese thesaurus.

References

1. Brill, Eric, A Simple rule-based Part of Speech tagger, In Proceedings of the third Conference on ANLP ACL, Trento, Italy, 1992.
2. Brown, P., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Lafferty J., Mercer R., Roosin P., A Statistical Approach to machine translation, Computational Linguistics, 16:2, pages 79-85, 1990.
3. Brown, P., Lai J., Mercer R. L., Word Sense Disambiguation using Statistical Methods, In proceedings of the 29th Annual ACL Meeting, page 264-270, 1991.
4. Brown, P., Stephen A. Della Pietra, Vincent J. Della Pietra, John D. Lafferty, and Robert L. Mercer, Analysis, Statistical Transfer, and Synthesis in Machine Translation, In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, page 83-100, 1992.
5. Brown, P. , Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, Vol. 19, No. 2, page 263-311, 1993.
6. Chang, Jyun-Sheng and Huey-Chyun Chen, Using Partially Aligned Parallel Text and Part-of-Speech Information in Word Alignment, In proceedings of the First Conference of the AMTA, page 16~23, 1994.
7. Chao, Yuen Ren, A Grammar of spoken Chinese, University of California Press, 1968.
8. Dagan, Ido, Kenneth W. Church and William A. Gale, Robust Bilingual Word Alignment for Machine Aided Translation, In Proceedings of the Workshop on Very Large Corpora : Academic and Industrial Perspectives, page 1-8, 1993.
9. Daille, Gaussier and Lange, Towards Automatic Extraction of Monolingual and Bilingual Terminology, In COLING-94, page 515-521, 1994.
10. Eijk, Pim Vander, Automating the Acquisition of Bilingual Terminology, In Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics, 1993.
11. Fujii, Hideo and W. Bruce Croft, A Comparison of Indexing Techniques for Japanese Text Retrieval, In proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval, page 237-246, 1993.
12. Gale, W.A. and K.W. Church. Identifying Word Correspondences in Parallel Texts, In Proceedings of the Fourth DARPA Speech and Natural Language Workshop, pages 152-157, Pacific Grove, CA., February 1991.
13. Kupiec, Juilian, An algorithm for finding noun phrase correspondences in bilingual corpora, In Proceeding of the 31st Annual Meeting of the ACL, 1993.
14. Longman, Longman English-Chinese Dictionary of Contemporary English, Published by Longman Group (Far East) Ltd., Hong Kong, 1992.
15. Matsumoto, Y. et al. Structural Matching of Parallet Texts, Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 1-30, Ohio, USA, 1993.
16. McArthur, T. Longman Lexicon of Contemporary English, Published by Longman Group (Far East) Ltd., Hong Kong, 1992.
17. Mei, J.J. et al., Tongyici Cilin (Word Forest of Synonyms), Tong Hua Publishing, Taipei, 1993 (traditional Chinese edition of a simplified Chinese edition published in 1984).